(2)

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 16669.25-MA	2. GOVT ACCESSION NO.: AD-A119 018 / N/A
3. RECIPIENT'S CATALOG NUMBER: N/A	
4. TITLE (and Subtitle): Using Biweight M-Lstimates in the Two-Sample Problem. Part 1: Symmetric Populations	5. TYPE OF REPORT & PERIOD COVERED: Reprint
	6. PERFORMING ORG. REPORT NUMBER: N/A
7. AUTHOR(s): Karen Kafadar	8. CONTRACT OR GRANT NUMBER(s): DAAG29 79 C 0205
9. PERFORMING ORGANIZATION NAME AND ADDRESS: National Bureau of Standards, Washington, DC 20234	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: N/A
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, P. O. Box 12211, Research Triangle Park, NC 27709	12. REPORT DATE: 1982
	13. NUMBER OF PAGES: 19
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Submitted for announcement only.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

COMMUN. STATIST.-THEOR. METH., 11(17), 1883-1901 (1982)

USING BIWEIGHT M-ESTIMATES IN THE TWO-SAMPLE PROBLEM PART 1: SYMMETRIC POPULATIONS

Karen Kafadar

Statistical Engineering Division
National Bureau of Standards

Key Words and Phrases: Student's t statistic; Monte Carlo simulation; robust confidence intervals; robustness of efficiency; robustness of validity.

ABSTRACT

We propose replacing the usual Student's-t statistic, which tests for equality of means of two distributions and is used to construct a confidence interval for the difference, by a biweight-"t" statistic. The biweight-"t" is a ratio of the difference of the biweight estimates of location from the two samples to an estimate of the standard error of this difference. Three forms of the denominator are evaluated: weighted variance estimates using both pooled and unpooled scale estimates, and unweighted variance estimates using an unpooled scale estimate. Monte Carlo simulations reveal that resulting confidence intervals are highly efficient on moderate sample sizes, and that nominal levels are nearly attained, even when considering extreme percentage points.

1. INTRODUCTION

The use of Student's t in constructing confidence intervals for the difference in location of two populations is a common practice. It is well known that this procedure is uniformly most powerful unbiased when the underlying populations follow Gaussian

distributions with the same variance (Lehmann 1959). When the distributions are in fact even slightly stretched-tailed, however, studies show that, while the Student's t interval nearly maintains its validity under the null hypothesis (Yuen and Dixon 1973, Lee and D'Agostino 1976), the power may be substantially reduced (Yuen and Dixon 1973). (More recently, see Benjamini 1980 for conditions under which one-sample Student's t is conservative.) In order to achieve "robustness of efficiency" in addition to "robustness of validity" (as defined in Tukey and McLaughlin 1963), this study proposes the use of biweights in a two-sample "t"-like statistic, which we shall call biweight-"t". The two-sample problem raises the issues of combining information on scale of the data and on variance of the numerator of biweight-"t". We shall attempt to judge when such borrowing of scale information may be justified. This report concentrates on small to moderate sizes of samples from symmetric populations; the unsymmetric case is treated in a forthcoming paper. Section 2 deals with the case of equal sample sizes. Section 3 considers unequal sample sizes, for which variance estimates may be weighted by their sample sizes. Section 4 examines the performance of biweight-"t" when the samples have different scales. A brief comparison of biweight-"t" intervals with other familiar procedures is made in Section 5, and Section 6 concludes with an example and strategies for the two-sample case.

2. EQUAL SAMPLE SIZES.

2.1 Form of two-sample biweight-"t" and concepts.

Let $x_{11},\ldots,x_{n_j,j} \sim F_j((x-\mu_j)/\sigma_j)$, $j = 1,2$, denote samples from two symmetric populations. Then the two-sample biweight-"t" takes the form:

$$"t"_{bi} = (T_1-T_2)/S$$

where each T_j is a biweight estimate of location and the squared denominator estimates the variance of the numerator:

$$S^2 = \hat{Var}\,(T_1-T_2).$$

For a definition of the biweight and its associated variance, the reader is referred to Mosteller and Tukey (1977). For a single

sample, $y_1,\ldots,y_n$, the only major difference between their calculation of the biweight estimate of location and that used here is in the choice of scale: 6·MAD (median absolute deviation) has been replaced by $(6\cdot s_{bi})$, where

$$\{\tilde{u}_k\} = (\tilde{u}_1, \ldots, \tilde{u}_n) = \{(y_k-\text{median})/9\cdot\text{MAD}\}$$

$$s_{bi}^2 = n\cdot q(\{\tilde{u}_k\})$$

$$q(\{\tilde{u}_k\}) = \sum_{k=1}^{n} \psi^2(\tilde{u}_k)/\{[\sum_{k=1}^{n} \psi'(\tilde{u}_k)]\cdot\max\{1, -1+\sum_{k=1}^{n} \psi'(\tilde{u}_k)]\} \quad (1)$$

and the psi function is given by

$$\psi(u) = u(1-u^2)^2 = u\cdot w(u) , \quad |u| \le 1$$
$$= 0 , \quad \text{else.}$$

One then solves for T iteratively via the equation

$$T^{(h)} = \sum_{k=1}^{n} y_k w(u_k)/\sum_{k=1}^{n} w(u_k), \quad u_k = [y_k - T^{(h-1)})/(6\cdot s_{bi})]. \quad (2)$$

The iteration starts with the median and ceases when the change is less than one part in the fourth decimal place. An estimate of the variance of T may be obtained from a finite-sample approximation to the theoretical asymptotic variance (cf. Huber 1981, p. 45):

$$S_T^2 = \hat{\text{Var}}(T) = q(\{u_k\}) \quad (3)$$

where the $\{u_k\}$ are defined in (2). (The motivation for these changes is discussed in Kafadar 1981, henceforth referred to as [K81].)

When we have two samples, we compute T and S_T for each sample. If we denote these by T_j and S_j (j=1,2), our two-sample biweight-"t" statistic then takes the form

$$\text{"t"} = (T_1 - T_2)/(S_1^2 + S_2^2)^{1/2} . \quad (4)$$

The variance estimates S_j^2 will be weighted by sample size in Section 3. In the remaining sections, we will drop the subscript on "t"$_{bi}$, as the form of "t" will always involve the biweight estimates as defined above.

2.2 Evaluation criteria.

Performance of biweight-"t" will be evaluated on three different distributions:

- Gaussian

- One-Wild (n-1 observations from N(0,1).
  1 unidentified observation from N(0,100))
- Slash (N(0,1) deviate / independent Uniform[0,1] deviate).

These three situations are likely to cover a reasonably broad range of stretched-tailed behavior (Rogers and Tukey 1972).

Robustness of efficiency may be evaluated in several ways. In this study, the success of biweight-"t" will be measured primarily in terms of "efficiency" of the expected confidence interval length (ECIL), i.e.,

$$\text{eff}(\alpha) = [\text{ECIL}_{\text{min}}(\alpha)/\text{ECIL}_{\text{actual}}(\alpha)]^2$$

where $\text{ECIL}_{\text{actual}}(\alpha)$ was defined by Gross (1976) as

$$\text{ECIL}_{\text{actual}}(\alpha) = 2(\alpha/2 \text{ \%-point of "t")} \cdot \text{ave(denominator of "t")},$$

and $\text{ECIL}_{\text{min}}(\alpha)$ is the shortest confidence interval we could expect for the given situation at hand. For the Gaussian, these are, of course, Student's t intervals, an approximation, derived in [K81], is used for $\text{ECIL}_{\text{min}}(\alpha)$ in the One-Wild and Slash situations.

Furthermore, for practical ease of use, we wish to approximate the distribution of biweight-"t" by one from a standard family of distributions. The most likely candidate here is Student's t, with some chosen number of degrees of freedom. This chosen number may be determined by comparing the calculated percent point of "t" to a Student's t table; i.e., the matching of ("t" critical point, α) to (degrees of freedom). The critical points of the distribution were all computed via a Monte Carlo swindle, the details of which may be found in Kafadar (1979). The sets of samples were those used in the Princeton Robustness Study (Andrews et al. 1972), each simulated situation involved either 640 or 1000 samples of sizes 5, 10, and 20.

2.3 Asymptotic Distribution of "t".

That "t" of (4) has an asymptotic Gaussian distribution is clear by the following argument: for the jth population,

$$n^{1/2}(T_j - \mu_j) \xrightarrow{D} N[0, E_j\Psi^2/(E_j\Psi')^2] ,$$

where the subscript of E denotes the distribution; e.g.,

$$E_1\Psi^2 = \int \Psi^2[(x-T_1)/(cs_1)]dF_1(x) \qquad (5)$$

for an arbitrary constant c (e.g., c=6 in (2)). Hence,

$$n^{1/2}[(T_1-T_2)-(\mu_1-\mu_2)] \xrightarrow{D} N[0, \sum_{j=1}^{2} E_j\Psi^2/(E_j\Psi')^2] .$$

Since (cf. Carroll 1978)

$$n \cdot S_j^2 \xrightarrow{D} E_j\Psi^2/(E_j\Psi')^2$$

we have by Slutsky's theorem that

$$[T_1-T_2)-(\mu_1-\mu_2)]/(S_1^2+S_2^2)^{1/2} \xrightarrow{D} N(0,1). \qquad (6)$$

2.4 Borrowing Scales.

Since each of the biweights in the numerator and each of the variance estimates in the denominator of "t" requires an estimate of scale, we may consider a pooled estimate if we believe that both samples have common scale. As shown in [K81], such a pooled estimate in the one-sample "t" can substantially reduce the variability in our results.

Table I gives the results of two-sample "t" when both samples

Table I
Biweight-"t" with pooled scales ($F_1 = F_2$)

tail area	Gaussian			One-Wild			Slash		
	%-pt	df	eff	%-pt	df	eff	%-pt	df	eff
A) $n_1=n_2=20$									
.05	1.663	71.1	97.4	1.662	91.0	95.0	1.677	47.8	69.4
.025	2.002	57.9	97.3	1.996	67.2	95.0	2.004	54.7	71.2
.001	3.279	43.9	96.5	3.290	43.3	94.3	3.228	62.2	78.6
.0001	4.080	41.9	96.9	4.111	38.7	93.3	3.984	55.8	81.6
.00001	4.813	41.4	97.1	4.894	36.8	91.4	4.720	49.5	82.1
B) $n_1=n_2=10$									
.05	1.692	33.3	93.7	1.693	32.7	86.6	1.700	28.4	70.9
.025	2.053	26.8	93.5	2.052	27.0	86.9	2.020	40.8	75.5
.001	3.537	20.7	93.0	3.571	19.3	86.7	3.277	46.5	92.7
.0001	4.546	19.9	93.4	5.054	16.9	84.0	4.341	33.6	99.8
.00001	5.581	19.6	93.8	5.955	16.0	81.5	4.923	35.4	101.5
C) $n_1=n_2=5$									
.05	1.849	8.4	91.1	1.769	13.2	60.5	1.790	11.4	68.1
.025	2.348	7.3	86.9	2.248	9.4	59.4	2.269	8.8	77.7
.001	7.267	4.0	34.5	6.483	4.5	32.3	7.927	3.7	113.1
.0001	16.658	3.6	13.5	16.326	8.7	12.1	28.573	2.9	47.0
.00001	25.061	3.9	11.4	22.755	4.1	14.0	66.471	2.7	1.8

have the same size and underlying distribution. (Additional percent points are available from the author.) Both biweights in the numerator have been scaled by s_{bor}, where

$$s_{bor} = [(n_1+n_2)\cdot q(\{u_{i1}, u_{i2}\})]^{1/2} \qquad (7)$$

$$u_{ij} = (x_{ij}-T_j)/(9\cdot s_j^{(0)})$$

$$s_j^{(0)} = \underset{I}{med}|x_{ij}-T_j^{(0)}|, \quad T_j^{(0)} = \underset{I}{med}\, x_{ij}$$

The subscript refers to a scale estimate which "borrows" width information from more than one sample.

Table I reveals extremely high performance for $n_i \geq 10$. In particular, the resulting confidence intervals for the Gaussian are trivially less efficient than if we knew the true underlying distribution (93% or higher) and are seldom more than 20% wider than the minimum ECIL for any situation. Furthermore, we are entitled to the full degrees of freedom in our approximation to a Student's t distribution, across a broad range of α-levels.

To be conservative, we might wish to approximate "t" by a Student's t on nine-tenths of the nominal degrees of freedom (ndf = n_1+n_2-2). For $\alpha > .01$ and $n_i \geq 10$, the actual error rate is only slightly smaller than the nominal (no less than 85% of the nominal). As we go further into the tails, however, the actual error rates may be as low as 30% of the nominal (even lower for Slash, n=10). While the robustness of classical procedures for extreme α-levels has not been investigated, a comparison with the values in Lee and D'Agostino (1976) indicates that this procedure is highly robust of validity at $\alpha = .05$, presumably this robustness extends to the extreme α-levels as well.

2.5 Different distributions: separate scale estimates.

All three distributions in this study are derived from the Gaussian with unit variance. This fact, however, does not imply that a pooled scale is appropriate when our samples come from different populations, as Table II(A) reveals. When our two samples do not both have the same underlying distributional shape, ECIL efficiency is still high, but the equivalent degrees of

Table II
Biweight-"t" with pooled and unpooled scales ($F_1 \neq F_2$)

tail area	Gaussian, One-Wild			Gaussian, Slash			One-Wild, Slash		
	%-pt	df	eff	%-pt	df	eff	%-pt	df	eff
A) Pooled Scales									
1) $n_1=n_2=20$									
.05	1.665	76.1	96.1	1.734	17.9	91.6	1.725	20.0	87.6
.025	2.001	58.9	96.0	2.090	19.4	93.3	2.072	22.4	90.0
.001	3.303	41.0	94.9	3.523	21.3	96.4	3.456	24.7	94.7
.0001	4.125	37.5	93.7	4.468	21.7	96.8	4.376	24.3	95.3
.00001	4.921	35.5	92.3	5.389	22.3	96.7	5.298	24.0	94.5
2) $n_1=n_2=10$									
.05	1.772	12.9	99.0	1.759	14.3	92.0	1.749	15.6	84.4
.025	2.171	12.4	97.1	2.125	15.5	94.6	2.105	17.5	87.4
.001	3.902	12.3	89.7	3.695	15.8	98.5	3.631	17.4	92.5
.0001	5.176	12.6	85.2	4.943	14.4	91.7	4.895	14.9	84.8
.00001	6.557	12.7	81.2	6.343	13.7	82.3	6.361	13.7	74.2
3) $n_1=n_2=5$									
.05	2.213	3.5	54.7	2.163	3.8	62.7	1.975	5.5	50.8
.025	3.137	3.1	42.5	3.005	3.4	58.2	2.641	4.6	50.8
.001	38.305	1.8	1.2	25.437	2.0	9.1	11.178	2.9	31.6
.0001	84.351	2.0	0.5	96.043	2.0	1.0	28.161	2.9	5.8
.00001	130.382	2.3	0.5	166.625	2.4	0.5	48.053	3.0	3.7
B) Unpooled Scales									
1) $n_1=n_2=20$									
.05	1.672	56.7	95.7	1.703	27.2	72.9	1.694	32.1	72.6
.025	2.010	49.2	95.6	2.041	30.4	74.9	2.028	36.1	74.7
.001	3.316	38.6	94.5	3.368	31.9	80.9	3.327	37.1	81.7
.0001	4.142	36.1	93.3	4.251	29.3	81.9	4.157	34.9	84.4
.00001	4.940	34.7	91.9	5.145	27.5	81.3	4.962	33.7	86.1
2) $n_1=n_2=10$									
.05	1.785	11.8	97.1	1.768	13.3	73.5	1.760	14.2	71.7
.025	2.186	11.7	95.2	2.122	15.8	76.7	2.111	16.9	74.8
.001	3.924	12.1	88.2	3.681	16.1	80.1	3.662	16.6	78.2
.0001	5.196	12.4	84.1	4.937	14.5	74.2	4.958	14.3	71.0
.00001	6.574	12.6	80.4	6.349	13.4	66.3	6.439	13.3	62.2

freedom is low. This not surprising, for (7) is designed to estimate a common scale. A comparison of the distributions based on a different characteristic of width, such as a pseudo-variance quantity, shows that the Slash is considerably wider than either the Gaussian or the One-Wild (cf. Rogers and Tukey 1972).

The scale estimate (7) borrows from both samples and is used in four places in our "t"-statistic. In general, of course, we shall not know when we are entitled to borrow. More importantly, this pooled scale violates the independence assumption in the numerator. It is true that the asymptotic distribution (6) depends

only upon the consistency (not the dependence) of the scale estimates in the numerator (T_1-T_2). However, we shall be applying this result to relatively small sample sizes. While the dependence between numerator and demoninator did not affect the efficiency of a biweight-"t" in the one-sample problem (cf. [K81]), it is not clear how the increased dependence in the numerator of "t" will alter its distribution on finite sample sizes.

To illustrate the effect of eliminating this dependence between the variables in the numerator, Panel B of Table II shows the results based on unpooled scales. Curiously, despite the incompatibility of scales in the Gaussian-Slash and Gaussian-One Wild pairs, biweight-"t" with pooled scales gives higher ECIL efficiency but slightly less degrees of freedom when n_1=n_2=20. Overall, we could be fairly confident in a comparison of two-sample "t" to a Student's t on 0.9(ndf), if we knew when and when not to borrow.

One criterion on which to base a decision applies a weight function to the logarithms of the scale estimates. This is explored in Kafadar (1980); preliminary results on small sample sizes are encouraging. Although formal tests of equal variances are beyond the scope of this paper, one might decide to borrow on the basis of the relative sizes of s_{bi} for the two samples. In the absense of a formal test, overall we conclude that "t" based on pooled scales allows roughly .9(ndf) for all but the extreme α-levels, and roughly .8(ndf) for the unpooled case.

When n_1=n_2=5, degrees of freedom are substantially lower than the nominal 8, and ECIL efficiency is often below 50%, even for the Gaussian case, where the biweight typically performs well. An explanation for this is explored in [K81]: the occasional presence of one or more observations which receive zero weight will lead to misleading estimates of scale, thereby affecting the distribution of "t". For small samples, the distribution of "t" can be characterized much more usefully by conditioning on the values of the sum of the biweight weights. These conditional results will not be shown here but are available from the author.

3. UNEQUAL SAMPLE SIZES.

This case is treated separately, because of the dependence of the variance estimates on sample size in the denominator.

3.1 Asymptotic Distribution of analogous two-sample statistic.

If we believe that our biweights in the numerator have the same variance, a common assumption in the usual two-sample approach, we may wish to pool our variance estimates in a "borrowed" (via mean squares) denominator:

$$\hat{\mathrm{Var}}(T_1-T_2) = S_{bor}^2 = [(n_1+n_2-2)^{-1}\sum_{j=1}^{2} n_j(n_j-1)S_j^2](n_1^{-1}+n_2^{-1}). \quad (8)$$

A borrowed-"t" then takes the form:

$$\text{"t"}_{bor} = [(T_1-T_2)-(\mu_1-\mu_2)]/S_{bor} \ . \quad (9)$$

In computing T_j and S_j in (8) and (9), one may (or may not) choose to use a pooled scale estimate as in (7).

The denominator in (9) weights the estimated variances of the statistics in the numerator according to the sample size. Such an approach would not be reasonable if $\mathrm{Var}(T_1) \neq \mathrm{Var}(T_2)$. For such unequal variance cases, we consider separate estimates of the variance in an unborrowed denominator (cf. Welch's approach to the Behrens-Fisher problem, Welch 1938):

$$\text{"t"}_{unbor} = [(T_1-T_2)-(\mu_1-\mu_2)]/(S_1^2 + S_2^2)^{1/2} \ , \quad (10)$$

since the variance of the numerator may also be estimated by

$$S^2_{unbor} = S_1^2 + S_2^2 \ . \quad (11)$$

This distinction did not of course arise in Section 2, for then (8) and (10) reduce to the same formula.

That the two forms of two-sample "t" do indeed have asymptotic Gaussian distributions under the null hypothesis can be seen as follows. Following the lines of the argument in Section 2.3, we know that

$$U_i = \sqrt{n_i}(T_i-\mu_i)[E_i\Psi^2/(E_i\Psi')^2]^{-1/2} \xrightarrow{D} N(0,1), \ i = 1,2 \quad (12)$$

where the notation for the expectations is defined in (5). Furthermore, if $F_1=F_2$, then the denominators in (12) are the same for both samples, so

$$\hat{\sigma}^2(n_1,n_2) = [n_1(n_1-1)S_1^2+n_2(n_2-1)S_2^2]/(n_1+n_2-2) \xrightarrow{P} E\Psi^2/(E\Psi')^2.$$

Hence, we have that "t"$_{bor}$ may be written

$$\text{"t"}_{bor} = [(U_1/\sqrt{n_1} - U_2/\sqrt{n_2})(n_1^{-1}+n_2^{-1})^{-1/2}\cdot[\sqrt{E\Psi^2/(E\Psi')^2}/\hat{\sigma}(n_1,n_2)].$$

$$= [U_1\{1+(n_1/n_2)\}^{-1/2} - U_2\{1+(n_2/n_1)\}^{-1/2}]\cdot[\sqrt{E\Psi^2/(E\Psi')^2}/\hat{\sigma}(n_1,n_2)].$$

If $n_1 \to \infty$ and $n_2 \to \infty$ in such a way that $n_1/n_2 \to K < \infty$,

$$[1+(n_2/n_1)]^{-1/2} \to (1+K)^{-1/2}, \quad [1+(n_1/n_2)]^{-1/2} \to [K/(1+K)]^{1/2}.$$

Hence, using Slutsky's theorem in conjunction with the convergence in distribution in (12), we conclude that "t"$_{bor}$ has an asymptotic Gaussian distribution. If $F_1 \neq F_2$, then "t"$_{unbor}$ is appropriate, for which the proof is similar.

3.2 Borrowing versus unborrowing: scales and denominators.

When we no longer have equal sample sizes, we might be *cautious and prefer not to borrow estimates* of either scale or biweight "variance". We know that such a cautious procedure may be quite wasteful of valuable information, especially when one sample has only five observations. On the other hand, biweight variances need not be the same for all distributions, and unwarranted borrowing in such cases may prove misleading. In this section we investigate the effects of various borrowing possibilities.

For the sake of brevity and for ease of comparison, we shall limit our attention to the efficiency of biweight-"t" at $\alpha = .001$ as representative of the behavior of "t" over the range $.00001 \leq \alpha \leq .05$. Table III shows these results, where the denominator of "t" is:

A) S_{bor}, borrowed scales: "complete borrowing";

B) S_{bor}, unborrowed scales: "incomplete borrowing",

C) S_{unbor}, unborrowed scales: "complete unborrowing".

When the distributions are the same, there is nearly always advantage to complete borrowing, as seen most dramatically when both underlying densities are Gaussian. In these cases, we may again approximate the distribution of "t" by a Student's t with the

Table III
Matched degrees of freedom and ECIL efficiencies at α=.001 for two-sample biweight-"t": Unequal sample sizes(1)

		complete borrowing		incomplete borrowing		complete unborrowing		nominal
		df	eff	df	eff	df	eff	df
A)	$F_1 = F_2$ (2)							
	G 10 G 20	30.0	96.1	25.5	94.45	17.4	85.11	28
	W 10 W 20	29.4	92.0	24.3	88.18	16.8	76.08	28
	S 10 S 20	27.9	80.6	21.7	73.75	37.1	80.45	28
	G 5 G 10	14.5	95.7	10.6	85.78	6.4	58.24	13
	W 5 W 10	11.8	75.9	9.9	65.40	6.3	37.72	13
	S 5 S 10	13.7	384.6	10.9	330.41	5.5	211.06	13
	G 5 G 20	24.1	95.2	14.8	83.92	4.7	32.06	23
	W 5 W 20	14.2	79.0	12.5	70.84	4.7	19.98	23
	S 5 S 20	10.0	427.3	10.1	419.14	5.5	211.06	23
B)	$F_1 \neq F_2$							
	G 10 W 20	40.1	96.4	30.2	92.82	17.7	84.23	28
	G 10 S 20	∞	92.1	∞	82.20	36.3	92.64	28
	W 10 G 20	23.3	91.6	21.1	88.84	16.2	75.46	28
	W 10 S 20	∞	82.3	∞	78.43	40.6	88.57	28
	S 10 G 20	8.4	99.7	8.6	82.52	14.4	65.23	28
	S 10 W 20	9.0	99.9	9.1	83.94	16.9	68.84	28
	G 5 W 10	13.9	87.8	10.7	78.16	6.6	59.28	13
	G 5 S 10	59.0	77.1	53.2	64.57	11.6	75.53	13
	W 5 G 10	12.4	84.7	9.5	67.34	6.0	35.29	13
	W 5 S 10	19.2	50.4	28.1	53.78	9.8	52.83	13
	S 5 G 10	1.9	14.7	4.4	273.40	3.6	102.15	13
	S 5 W 10	1.9	11.5	4.6	267.82	3.7	105.87	13
	G 5 W 20	27.6	92.4	15.9	80.67	4.7	32.10	23
	G 5 S 20	∞	81.6	∞	72.50	12.8	74.29	23
	W 5 G 20	15.2	86.5	11.6	71.42	4.6	18.78	23
	W 5 S 20	∞	56.4	∞	56.30	11.8	48.96	23
	S 5 G 20	2.0	38.2	4.6	489.44	3.1	73.52	23
	S 5 W 20	2.0	44.2	4.6	467.52	3.1	74.78	23

(1) Standard errors for critical points from which degrees of freedom were matched and ECIL efficiencies were calculated fell in the range 0.028 to 0.331 for α = .001.

(2) F_j represents underlying distribution for sample j: G = Gaussian, W = One-Wild, S = Slash.

nominal degrees of freedom. When the distributions are the same, a conservative matching would be 0.9(ndf). When one distribution is Slash, incomplete borrowing appears slightly more successful.

Finally, we remark that there are some cases for which "t" in any of the three forms appears totally unsuccessful (e.g., n=5 Slash, with anything else). This is primarily due to the nature of small samples: there is a chance (about 5% in the Gaussian) that one or two bonafide observations will occur far enough away from the bulk of the data so as to be inappropriately downweighted by any reasonably robust procedure. When the smaller sample is

restricted to be such that the sum of the biweight weights is high, efficiencies on the biweight-"t" intervals are slightly higher than those in Panel B. A solution may well depend on an appropriate use of the weight distribution in these small samples.

4. UNEQUAL WIDTHS.

4.1 Unborrowed denominators.

When our samples have different scales, a Welch-like unborrowed denominator of the form (11) is a safe (but conservative) approach. To evaluate the performance of biweight-"t" in the presence of unequal widths, we multiply the observations of one of the distributions by either $\sqrt{2}$ or 2, yielding "variance" ratios between 2 and 4. A moderate difference in scales was chosen to provide some indication of the effect in practical applications.

In Table IV, we show some trials of "t"$_{unbor}$ either when $F_1 \neq F_2$, $n_1 = n_2$ or when $F_1 = F_2$, $n_1 \neq n_2$. (As in Table III, only the results for $\alpha = .001$ are shown.) Notice that our previous matching of the distribution to a Student's t on 0.8(ndf) for unpooled scales would be conservative. This is similar to the conservative nature of Welch's unborrowed t-statistic (e.g., as shown in Lee and D'Agostino 1976, Welch 1938). Approximating the distribution of "t"$_{unbor}$ by a Student's t on 0.9(ndf) instead, we see from Table IV that the actual levels are still often less than half the nominal. In terms of robustness of efficiency, however, ECIL efficiency typically exceeds 50%.

As a final comment on the interval problem for samples of varying widths, we mention the concept of transformation, a familiar data analytic tool in such situations. When comparing several batches of data, Tukey (1977, chapter 3,4) draws attention to the importance of choosing a re-expression of the data for which the amounts of spread are roughly the same across batches. Such re-expression may be useful in dealing with the "unequal variances" problem of this section. For example, Anscombe's (1948) variance stabilizing transformations of Poisson data have been shown to produce more similarity in spread. The results of biweight-"t"

Table
Matched degrees of freedom and ECIL efficiency for biweight-"t" at $\alpha = .001$: "σ_1" $\neq$ "σ_2"

F_1[1]	n_1	F_2	n_2	σ_2^2/σ_1^2	matched d.f.	ECIL eff.	actual α / nominal α [3]
A) $F_1 = F_2$							
G	10	G	20	2	∞[2]	90.90	.341
G	10	G	20	4	∞	92.46	.038
W	10	W	20	2	∞	82.94	.379
W	10	W	20	4	∞	85.33	.045
S	10	S	20	2	∞	262.61	.255
S	10	S	20	4	∞	208.25	.073
G	20	G	10	2	∞	72.46	.469
G	20	G	10	4	∞	56.07	.286
W	20	W	10	2	∞	62.26	.593
W	20	W	10	4	∞	47.62	.474
S	20	S	10	2	∞	224.64	.177
S	20	S	10	4	∞	155.83	.030
B) $F_1 \neq F_2$							
G	20	W	20	2	∞	91.21	.127
G	20	W	20	4	∞	85.55	.007
G	10	W	10	4	∞	47.00	.533
G	20	S	20	2	∞	57.22	.238
G	20	S	20	4	∞	35.84	.038
W	10	S	10	2	∞	57.20	.307
W	10	S	10	4	∞	35.60	.038

(1) F_j represents underlying distribution for sample j: G = Gaussian; W = One-Wild; S = Slash.

(2) Indicates that biweight-"t" distribution is shorter-tailed than Gaussian.

(3) Actual $\alpha = P["t"_{bi} > t_{.9(ndf)}(.001)]$; nominal $\alpha = .001$.

discussed in Sections 2 and 3 (perhaps even the completely borrowed "t") may than be applied successfully to such re-expressed data.

5. COMPARISON WITH CLASSICAL AND NONPARAMETRIC INTERVALS.

Many practicing statisticians are reluctant to compute robust estimators or are satisfied with distribution-free methods. Even among users of robust methods, there has been disagreement concerning the efficiency of the biweight over robust estimators. To compare the performance of biweight-"t" with Student's t, a nonparametric and a Huber-type "t" interval, Table V presents the results from a separate Monte Carlo study. For each run, 1000 Gaussian or One-Wild samples of size 5, 10, or 20 were generated. Subroutine HH from Andrews et al. (1972) computed the Huber location estimate, and its standard error was estimated via (3) but

where $\Psi(u)$ was replaced by

$$\Psi_H(u) = u \quad |u| \leq 1.5$$
$$= 1.5 \quad |u| > 1.5 .$$

Nonparametric intervals based on the Wilcoxon rank sum test are described in Lehmann (1975). Student's t, Huber-"t", and biweight -"t" all used completely unborrowed denominators.

Table V reveals that Student's t is highly inefficient when even one of the samples is mildly contaminated (One-Wild, n=20); biweight-"t" intervals dominate the nonparametric intervals (sometimes by as much as 40%) as well as the Huber-"t" intervals. A constant of c=4 was also run for the biweight; efficiencies for

Table V
ECIL efficiencies for five different "t"-confidence intervals

	Student's t	Wilcoxon	Huber k=1.5	Biweight c=6	Biweight c=4
G 20 W 20					
α=.05	61.6	93.3	87.0	94.6	91.9
α=.001	61.5	90.0	75.8	90.1	84.9
α=.00005	58.8	86.0	68.4	86.5	79.2
W 10 W 20					
α=.05	39.2	84.3	77.9	87.7	87.4
α=.001	41.2	60.3	64.3	80.0	73.8
α=.00005	39.6	34.1	58.3	75.0	64.2
G 10 W 20					
α=.05	67.1	91.0	82.5	92.5	88.9
α=.001	65.5	84.8	69.3	82.2	75.3
α=.00005	62.0	68.2	64.2	74.3	67.9
G 10 W 10					
α=.05	48.9	86.0	80.1	89.8	88.1
α=.001	49.0	36.5	63.3	81.8	76.1
α=.00005	46.9	36.6	56.1	75.4	69.0
G 5 W 10					
α=.05	54.6	79.4	71.5	84.2	79.9
α=.001	52.5	36.7	51.9	66.6	60.0
α=.00005	49.0	-	45.4	61.3	53.7
G 5 W 5					
α=.05	36.9	27.8	62.3	68.5	72.1
α=.001	33.1	-	42.8	56.6	49.2
α=.00005	30.2	-	39.9	53.4	43.0

moderate contamination (≤10%) are only slightly lower than when c=6. The main message is that a robust "t" interval can lead to large gains in efficiency in long-tailed, symmetric situations.

6. AN APPLICATION AND CONCLUSIONS.

6.1 An Example for borrowed and unborrowed "t" intervals.

To gain some familiarity with the effect of borrowing scales on biweight comfidence intervals, we calculate them for a set of chemical measurements taken at the National Bureau of Standards. These data consist of the concentrations of polychlorinated biphenyl (PCB) in a motor oil solution as determined by gas chromatography (in units of milligrams per kilogram of oil). Each sample includes ten peak-by-peak comparisons of the oil fraction chromatogram with the chromatogram of a known standard mixture. The box plots of the data from four ampoules of solution are shown in Figure 1. Notice that the underlying assumption of symmetry

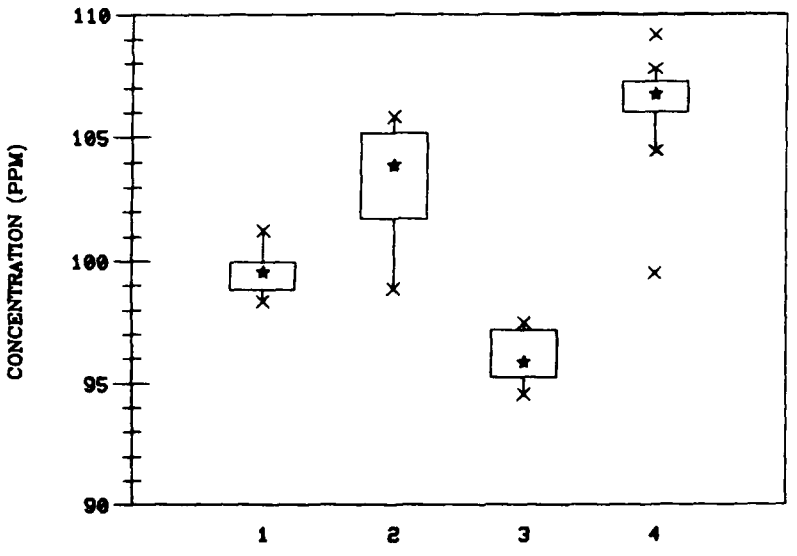

FIG. 1. Box Plots of Data from PCB's in oil.

does not seem unrealistic for these samples, and that some outliers are evident from ampoule 4.

While it appears that all four groups do not have a common scale, one might reasonably borrow scales between batches 1 and 4. If we are interested in all 6 pair-wise comparisons at the 95% level of confidence, each interval should be based on the 2.5%/6 ≈ .4%-point of the "t" distribution (.9x18=16.2 d.f. for pooled scales, .8x18 = 14.4 d.f. for unpooled scales). The pooled scale (7) between ampoules 1 and 4 is .988, from which biweights and associated variance estimates may be calculated to give a confidence interval of the form

$$(T_1 - T_4) \pm t_{16.2}(.004)(S_1^2 + S_4^2)^{1/2}$$
$$= (99.468 - 106.858) \pm (3.028)(.0706 + .176)^{1/2}$$
$$= -7.390 \pm 1.504 = (-8.894, -5.886).$$

(The corresponding Student's t interval, (-9.220, -4.026), is 1.7 times wider.) An unborrowed confidence interval for the difference between ampoules 1 and 2 is

$$(99.468 - 103.357) \pm t_{14.4}(.004)(.0706 + .587)^{1/2}$$
$$= -3.889 \pm 2.493 = (-6.382, -1.396).$$

(Welch's (1949) unborrowed confidence interval, using the formula for degrees of freedom on p. 295, is only trivially longer.) Comparing ampoules 2 and 4 gives a confidence interval of the form

$$(103.357 - 106.826) \pm t_{14.4}(.004)(.587 + .213)^{1/2}$$
$$= -3.469 \pm 2.749 = (-6.218, -0.720).$$

This last comparison illustrates the greater power in this procedure over the classical Student's t interval (-6.203, 0.423), which would not reject the hypothesis of a difference. (Had one used a Welch interval, since the equal-variance hypothesis rejects at level .10, it would have been even wider.)

These intervals do not represent the final data summary because additional information on the measurement process permits more accuracy in determining standard errors. For illustrative purposes, however, this information has been neglected.

6.2 Concluding comments for the two-sample case.

This study investigated the performance of a two-sample "t"

statistic when classical sample means and variance are replaced by their biweight counterparts. Although computationally more difficult than Student's t, the popular use of computers makes this disadvantage irrelevant. The primary advantage is that its distribution can be well approximated by one from the Student's t family, from which valid, yet efficient, confidence intervals for the difference in centers can be made.

Appropriate scaling for biweight-"t" can be important. We can choose to either pool estimates of scale (a wise move if in fact we have common underlying situations), or use separate estimates (slightly safer in cases of doubt). The distribution may be matched to Student's t on .9(ndf) (out to .1%-point) in the former case or .8(ndf) in the latter. In either case, the efficiency of the procedure (in terms of relative length of the interval) is upwards of 70%. The same applies when $n_1 \neq n_2$, if we weight the variance estimates proportional to their sample sizes ("borrowed" denominator). Small samples sizes (n=5) pose a problem only when the underlying population is extremely heavy-tailed (e.g., Slash).

A few trials of unborrowed denominators were run in situations where the samples did not have common width. For the most part, the 0.8(ndf) matching is quite conservative; .9(ndf) could be safely recommended for all but perhaps the most extreme percent points (.01% and beyond). When the underlying situations have the same width, we have better than 60% efficiency out to the .5% point. When the situations are different (either in distribution or in width) , the efficiency decreases with the increased difference in the distribution (in terms of the "heaviness" of the tails).

While further insight into the nature of the weight distribution may sugggest refinements, present results indicate that we may feel confident in constructing two-sample biweight-"t" intervals using tabulated Student's t percent points as outlined above. A subsequent report will investigate the performance of biweight-"t" when the underlying populations are unsymmetric.

ACKNOWLEDGEMENT

This report is based on sections in the author's Ph.D. dissertation (Princeton University, 1979). Research was supported in part by a contract with the U. S. Army Research Office, No. DAAG 29-76-0298, awarded to the Department of Statistics, Princeton University, Princeton, New Jersey. The author gratefully acknowledges Professors J. W. Tukey and P. Bloomfield for much helpful advice during the preparation and for numerous comments on early drafts of this paper, and Dr. Stephen N. Chesler of the National Bureau of Standards for providing the data in Section 6.1.

BIBLIOGRAPHY

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972). Robust Estimates of Location: Survey and Advances. Princeton University Press: Princeton, New Jersey.

Anscombe, F.J. (1948). The transformation of Poisson, binomial and negative binomial data. Biometrika 35, 246-254.

Benjamini, Yoav (1980). The behavior of the t-test when the parent distribution is long-tailed. Ph.D. Dissertation, Princeton University, Princeton, New Jersey.

Carroll, Raymond J. (1978). On almost-sure expansions for M-estimates. Ann. Statist. 6, No. 2, 314-318.

Gross, A.M. (1976). Confidence interval robustness with long-tailed symmetric distributions. J. Amer. Statist. Assoc. 71, 409-417.

Huber, Peter (1981). Robust Statistics. Wiley: New York

Kafadar, Karen (1981) ([K81]). A biweight approach to the one-sample problem. To appear in J. Amer. Statist. Assoc.

_____ (1979). A two-sample Monte Carlo swindle. Technical Report No. 153, Dept. of Statistics, Princeton University, Princeton, N.J.

_____ (1980). An empirical investigation of small samples from symmetric populations for constricting robust confidence intervals Technical Report No. 74, Dept. of Statistics, Oregon State University, Corvallis, OR.

Lax, David (1975). An interim report of a Monte Carlo study of robust estimates of width. Technical Report No. 93, Dept. of Statistics, Princeton University, Princeton, N.J.

Lee, Austin F.S and D'Agostino, Ralph B. (1976). Levels of significance of some two-sample tests when observations are from compound normals. Communications in Statistics A5, No. 4, 325-342.

Lehmann, E.L. (1959). Testing Statistical Hypotheses. Wiley: New York.

(1975) Nonparametrics: Statistical Methods based on ranks. Holden-Day: San Francisco.

Mosteller, F. and Tukey, J.W. (1977). Data Analysis and Regression: A second course in statistics. Addison-Wesley: Reading, MA.

Rogers, W.H. and Tukey, J.W. (1972). Understanding some long-tailed symmetrical distributions. Statistica Neerlandica 26, No. 3, 211-226.

Tukey, J.W. (1977). Exploratory Data Analysis. Addison-Wesley: Reading, MA.

Tukey, J.W. and McLaughlin, Donald H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. Sankhya, Series A, 25, 331-352.

Welch, B.F. (1938). The significance of the difference between two means when the population variances are unequal. Biometrika 29, 350-362.

(1949). Appendix to A. A. Aspin's tables. Biometrika 36, 293-6.

Yuen, Karen K. (1974). The two-sample trimmed t for unequal population variances. Biometrika 61, Vol. 1, 165-169.

Yuen, Karen K. and Dixon, W.J. (1973). The approximate behavior and performance of the two-sample trimmed t. Biometrika 60, Vol. 2, 369-374.

Received May, 1981; Revised April, 1982.

Recommended by Wm. H. Rogers, The Rand Corporation, Santa Monica, CA